DOCUMENT RESUME

ED 255 542 TM 850 119

AUTHOR Wilson, Kenneth M.
TITLE The Relationship of GRE General Test Item-Type Part Scores to Undergraduate Grades.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RR-84-38; GREB-81-22P
PUB DATE Feb 85
NOTE 52p.; Contains small print in some tables and figures.
PUB TYPE Reports - Research/Technical (143)

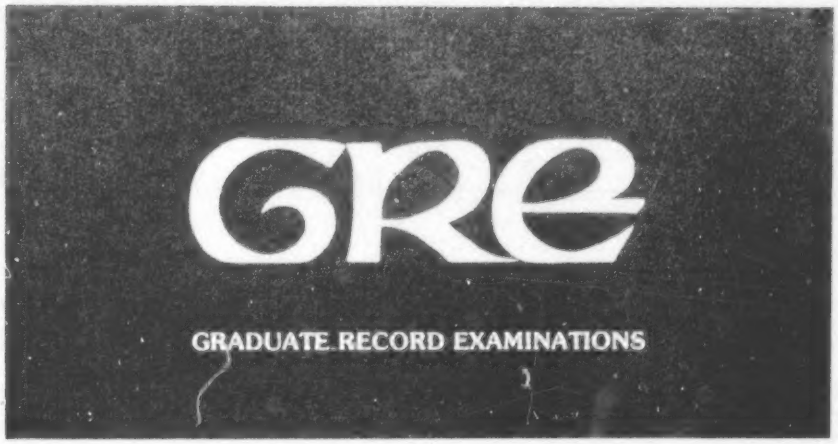
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *College Entrance Examinations; *Grade Point Average; Higher Education; Item Analysis; Majors (Students); *Predictive Validity; Quantitative Tests; Reading Comprehension; *Test Items; Undergraduate Students; Verbal Tests; Vocabulary Skills
IDENTIFIERS Analytical Tests; *Graduate Record Examinations

ABSTRACT

This Graduate Record Examination (GRE) study assesses: (1) the relative contribution of a vocabulary score (consisting of GRE General Test antonyms and analogies) and a reading comprehension score (consisting of GRE sentence completion and reading comprehension sets) to the prediction of self-reported undergraduate grade point average (GPA); and (2) criterion-related validity patterns for item-type part scores on the GRE quantitative and analytical measures. Data from GRE files for 9,375 examinees in 12 fields of study representing 437 undergraduate departments from 149 colleges and universities were standardized within each undergraduate department, and then pooled for analysis by field. There were differences by major field in average performance on the various item-type part scores within each test. The reading comprehension subtest carried most of the predictive load in the GRE verbal measure. Item-type part scores on the other measures also exhibited differential patterns of relationships with the self-reported undergraduate grade point average. The findings suggest that the different item types within the respective broad ability measures may be tapping somewhat unique skills and abilities and that further exploration of their potential contribution is in order. (Author/BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED255542



"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. Wedenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

THE RELATIONSHIP OF GRE GENERAL TEST ITEM-TYPE PART SCORES TO UNDERGRADUATE GRADES

Kenneth M. Wilson

GRE Board Professional Report GREB No. 81-22P
ETS Research Report 84-38

February, 1985

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

TM 850 119

Abstract

This study was undertaken (a) to assess the relative contribution of a vocabulary score made up of GRE General Test antonyms and analogies and a reading comprehension score made up of GRE sentence completions and reading comprehension sets to prediction of an academic criterion (self-reported undergraduate grade point average) and (b) to assess patterns of criterion-related validity for item-type part scores on the GRE quantitative and analytical measures as well.

The study was based on data from GRE files for 9,375 examinees in 12 fields of study representing 437 undergraduate departments from 149 colleges and universities. All data were standardized within each undergraduate department and then pooled for analysis by field.

There were differences by major field in average performance on the various item-type part scores within each test. The reading comprehension subtest was found to carry most of the predictive load in the GRE verbal measure (consistent with findings for the reading comprehension subscore on the SAT verbal measure). Item-type part scores on the other measures also exhibited differential patterns of relationships with the self-reported undergraduate grade point average.

The findings suggest that the different item types within the respective broad ability measures may be tapping somewhat unique skills and abilities and that further exploration of their potential contribution is in order.

Acknowledgments

To the Graduate Record Examinations Board under whose auspices this study was conducted;

To Richard H. Harrison for assistance in data management and analysis;

To Robert Altman, Richard Duran, Donald Powers, Donald Rock, and William B. Schrader for critical reviews of the original draft; and to Ruth Miller for editorial assistance in the preparation of the final draft.

These contributions are gratefully acknowledged.

Table of Contents

	Page
Acknowledgments	1
Contents	iii
Introduction	1
Study Design, Sample, and Procedures	3
Study Sample and Procedures	4
GRE Item-Type Part Score Data	6
Study Procedures	9
Major-Field Differences in Average Performance on GRE Item-Type Subtests	10
Exploratory Evaluation of Part Score Validity	15
The Verbal Test Part-Score Analysis	16
The Quantitative Test Part-Score Analysis	21
The Analytical Test Part-Score Analysis	23
Verbal, Quantitative, and Analytical Part Scores as a Battery	28
Comparability of Regression Results for Unequated and Equated Total Scores	32
Summary of Trends in Findings	32
Discussion	34
References	37
Appendix A. Supplementary Data on GRE Part Scores	39
Appendix B. Comparability of Part-Score Validity Profiles for Single Form and Multiple Form Unequated Score Samples	41
Appendix C. Factors Involved in the Use of Total vs Pooled Within-Group Correlations in Validation Research	43

The Relationship of GRE General Test Item-Type Part Scores to Undergraduate Grades

Kenneth M. Wilson
Educational Testing Service

Introduction

The GRE General (Aptitude) Test provides measures of developed verbal, quantitative, and analytical abilities.* Only total verbal, quantitative, and analytical scores are reported. However, the three measures include different types of items that are thought of as being different methods of measuring their respective constructs (Rock, Werts, & Grandy, 1982).

The verbal measure employs four types of questions or items: antonyms, analogies, sentence completions, and reading comprehension sets designed to test the ability to identify (a) words that are opposite in meaning, (b) words or phrases that are related to each other in the same way as other words or phrases, and (c) words that are logically and stylistically consistent with the sentence in which they appear; and (d) the ability to recognize in a reading passage the main ideas, information explicitly provided, implied ideas, the attitude of the author, and the like.

Three item types are employed in the quantitative measure: quantitative comparisons (testing the ability to reason quickly and accurately regarding the relative sizes of two quantities or to perceive that not enough information is available to make such a decision); discrete quantitative items measuring basic mathematical skills or regular mathematics (balanced among question requiring arithmetic, algebra, and geometry and designed to test basic mathematical skills and understandings of concepts, at levels applicable to individuals who have not specialized in mathematics); and data interpretation (testing the ability to synthesize information presented in tabular or graphic form, to select data appropriate for answering a question, and so on).

The 1981 revision of the analytical measure includes two item types: analytical reasoning items (testing the ability to understand a given structure of arbitrary relationships among fictitious entities, deduce new information from given relationships, and the like); and logical reasoning items (testing the ability to understand, analyze, and evaluate arguments, recognize the point of an argument or the assumptions on which it is based, analyze evidence, and the like).

Although a continuing effort is made to obtain empirical evidence regarding the validity of the total verbal, quantitative, and analytical

*For detailed descriptions of tests and item types, see, for example, ETS (1981). In October 1977, a restructured version of the GRE General Test including a newly developed analytical ability measure was introduced. Evidence of its predictive validity with respect to graduate grades was obtained in a cooperative study (Wilson, 1982). However, internal research indicated the need for some change in the item content of the 1977 analytical measure and, in October 1981, a revised analytical measure was introduced. See Wild, Swinton, and Wallmark (1982) for a review of factors involved in the 1981 revision.

scores for predicting performance in graduate study, little attention has been given to study of the predictive validity and diagnostic potential of part scores based on the various item types--in large part because of the lack of any compelling a priori evidential or theoretical basis for expecting differential predictive validity for part scores based on different item types measuring more general basic constructs such as verbal or quantitative ability.

For example, items regardless of type are selected on the basis of internal consistency criteria designed among other things to assure the comparative homogeneity of the respective ability measures. This is conducive to relatively high intercorrelations among items and between individual items and the total scores on the respective tests. Such conditions theoretically militate against the likelihood, for example, that predictions based on regression-weighted composites of part scores would be consistently better than predictions based on the total score (in which the potential item-type part scores are weighted roughly according to their length). Although factor analytic studies (for example, Powers & Swinton, 1981; Rock, Werts, & Grandy, 1982) have suggested that word knowledge (vocabulary) and reading items (reading comprehension) are distinguishable factorially, this evidence alone has not been sufficiently persuasive to suggest that predictions based on the "vocabulary" items and predictions based on "reading comprehension" items would be very different.

However, the need for an empirical evaluation of the predictive validity of item-type part scores on the GRE General Test was indicated by the results of undergraduate-level validity studies involving verbal item-type part scores on the College Board Scholastic Aptitude Test (SAT). For several years, vocabulary (VO) and reading comprehension (RC) scores have been reported in addition to the total SAT verbal score. The vocabulary score is based on antonyms and analogies and the reading comprehension score on sentence completions and reading comprehension sets. These items are completely parallel in type to those included in the GRE verbal measure.

Based on internal analyses of the results of 110 studies conducted by the College Board Validity Study Service (VSS) at ETS (Ramist, 1981a; 1981b) in which colleges had specified vocabulary, reading comprehension, and total SAT verbal scores as predictors of freshman grades, the following findings emerged:

- The average validity of the reading comprehension score alone (.373) was only .003 points lower than that for the entire verbal score (.376).

- In almost one-half of the samples studied, the observed validity of the reading comprehension score was actually greater than that for the SAT verbal score, including the vocabulary score, the validity of which was consistently lower than that of the reading comprehension score.

- When vocabulary and reading comprehension scores were combined in regression-weighted composites, the vocabulary score in a number of instances was negatively weighted, although its simple correlation with the GPA criterion was positive, indicating suppression of vocabulary

variance in reading comprehension--that is, suggesting that the criterion-related variance in the vocabulary measure was being tapped sufficiently by the reading comprehension measure with which the vocabulary score is substantially correlated.

- There was little improvement in predicting freshman grade point average when separate vocabulary and reading comprehension scores replaced the SAT total verbal score in regression equations including SAT mathematical scores and the high school record.

These results were inconsistent with expectation and raised questions regarding the relative predictive role of the SAT vocabulary and reading comprehension items.* The present study was undertaken to assess the relationship to academic performance of similarly constructed GRE vocabulary and reading comprehension item-type part scores (and of item-type part scores based on items in the quantitative and analytical tests as well).

Study Design, Sample, and Procedures

The academic performance criterion selected for this exploratory study was self-reported undergraduate grade point average (SR-UGPA) routinely supplied by most GRE examinees during the process of test-registration.** The SR-UGPA has been found to be a useful research surrogate for an officially computed UGPA as a predictor of graduate GPA (Wilson, 1982). Moreover, patterns of coefficients for GRE verbal, quantitative, and analytical scores vs SR-UGPA, computed for samples of undergraduate students majoring in selected fields (for example, Miller & Wild, 1979) appear to be similar to patterns of coefficients for these predictors vs graduate GPA (for example, Wilson, 1982).

It was reasoned that results of an exploratory study involving SR-UGPA as the academic performance criterion would provide a useful empirical basis for initial assessment of the validity of item-type part scores. Such a study would also contribute to further understanding of the utility of the SR-UGPA in research concerned with test validation.

---

*Several lines of inquiry have been initiated, including a study of the relationship of vocabulary and reading comprehension scores to self-reported high school rank, a study of the statistical properties of the four item-types included in the SAT verbal measure, and a study of the criterion-related validity of specific verbal item types on one form of the SAT verbal test (Schrader, 1984).

**Examinees are asked to report UGPA in the major field and UGPA over the last two college years. The criterion employed was the average of the two self-reported undergraduate grade point averages.

The study was designed to simulate conditions characteristic of graduate-level validity studies in which comparable data sets for several small departmental samples are pooled for analysis by field or discipline (for example, Wilson, 1979; 1982).

Study Sample and Data

The study sample and basic study data were taken from GRE files on examinees tested between Oct. 1, 1981, and Sept. 30, 1982. The study sample included only examinees who reported better communication in English than in any other language, who were tested as enrolled undergraduates or nonenrolled college graduates no more than two years beyond the bachelor's degree, and who named both a field of study and an undergraduate school. Following procedures described below, data were obtained for examinees representing both (a) a relatively large number of undergraduate departments from each of 10 to 15 fields representing a wide range of verbal vs quantitative emphasis (for example, engineering to English), with some fields of relatively mixed emphasis such as education and biology.

The records of examinees eligible for inclusion in the study (by enrollment, citizenship, language status, and data-availability criteria) were classified by reported undergradaute major field, and the fields were ordered in terms of the total number of designators. Within each field classification, examinees were distributed according to designated undergraduate school, and schools were ordered according to total number of designators without regard to field--that is, in terms of total volume of graduate-school bound, currently or recently enrolled students in the GRE pool.

The 20 most frequently designated fields are listed below, and those selected for the study are identified by asterisks:

psychology	political science*	economics*	computer science*
biology*	chemistry*	sociology*	other biosciences
English*	geology	mathematics*	other social sciences
nursing	business	music	physical education
education*	history*	electrical engineering*	agriculture*

English, history, sociology, and political science may be thought of as representing primarily verbal fields; chemistry, computer science, mathematics, electrical engineering, and economics were selected as representing primarily quantitative fields; and agriculture, biology, and education represent fields not clearly classifiable according to relative verbal and quantitative emphases.

Schools and departments wre selected, within each of the 12 field classifications, by specifying certain minimum Ns, set after inspection of the data, to lead to inclusion of 20 or more samples from undergraduate schools contributing varied numbers of students to the general GRE examinee pool. Results of the selection process are indicated in Table 1. Data on

Table 1

Distribution of Undergraduate Departmental Samples Included in the Study By Size and Field

Sample size	Number of undergraduate departmental samples by field												
	English[a]	History[b]	Sociology[b]	Polit Sci[a]	Chemistry[b]	Comp Sci[b]	Mathematics[c]	Elec Engin[a]	Economics[b]	Agriculture[d]	Biology[d]	Education[d]	All fields
100+										1		1	2
90-99										-		-	-
80-89										-		2	2
70-79										1		-	1
60-69										2		3	5
50-59	1									4	2	1	8
40-49	-	1		1				3		1	5	10	21
30-39	2	-		2		2		6	2	2	11	4	31
20-29	16	5	2	6	7	5		12	6	13	33	19	124
10-19	24	33	24	16	38	34	13	15	36	-	-	-	233
<10	-	-	-	-	-	-	10	-	-	-	-	-	10
No. of depts	43	39	26	25	45	41	23	36	44	24	51	40	437
No. of students	884	584	364	545	644	647	251	850	663	976	1318	1649	9375
Male(%)	34.2	54.8	25.8	57.2	67.2	69.6	62.5	88.3	62.9	59.7	45.9	12.0	49.4
Minority(%)	11.0	13.9	29.2	18.8	14.6	17.9	11.0	21.6	15.4	7.3	14.9	9.0	14.1

Note. An undergraduate departmental sample includes individuals naming a designated undergraduate major field and a designated undergraduate school who were taking the GRE General Test during 1981-82 as either (a) enrolled undergraduates or (b) nonenrolled bachelor's degree holders no more than two years beyond the bachelor's.

[a]Minimum N = 15; [b]Minimum N = 10; [c]Minimum N = 9; [d]Minimum N = 20

sex composition and minority representation in the sample, by field, are also shown.

As may be determined from Table 1, the study sample included 9,375 individuals from a total of 437 undergraduater departments in 149 different undergraduate institutions. In 8 of the 12 fields, the modal number of undergraduate majors per department was between 10 and 19, and distributions of Ns per department were positively skewed around these small modal values within each field. These conditions are quite similar to those encountered in graduate level validity studies.

GRE Item-Type Part Score Data

For each member of the study sample, operational GRE scaled verbal, quantitative, and analytical scores and corresponding item response data were available, based on one of six different forms of the GRE General Test that were used during 1981-82. Each form included the same total number of items, and the same number of items by type, as indicated below:

Variable	No. of items
Verbal Test	(76)
Antonyms	22
Analogies	18
Sentence completions	14
Reading passages	22
Quantitative Test	(60)
Quantitative comparison	30
Regular mathematics	20
Data interpretation	10
Analytical Test	(50)
Analytical reasoning	38
Logical reasoning	12

Raw total scores (based on the 76 verbal, 60 quantitative, and 50 analytical test items) were computed for each member of the study sample taking each form of the test, and raw part scores were computed for each of the nine item types indicated above; in addition, a vocabulary score based on the 40 antonyms and analogies items and a reading comprehension score based on the 36 sentence completions and reading passage sets were computed for each individual. All raw scores were computed using the total number right scoring procedures introduced during 1981-82.

The part scores are of differing lengths, with corresponding differences in reliability. For example, based on internal analyses of two forms of the GRE General Test administered during 1981-82 (Wallmark, 1982a; 1982b),

typical levels of reliability (estimated by Kuder-Richardson Formula 20) of the various GRE scores in general smaples of GRE examinees are approximately as follows:

Test	Typical form reliability
Verbal Test (Total)	.90+ (76 items)
Antonyms + Analogies + Sentence completions	.90 (54 items)*
Reading comprehension	.80+ (22 items)
Quantitative Test (Total)	.90 (60 items)
Quantitative comparison	.80+ (30 items)
Regular mathematics	.75+ (20 items)
Data interpretation	.60+ (10 items)
Analytical Test (Total)	.85+ (50 items)
Analytical reasoning	.80+ (38 items)
Logical reasoning	.60+ (12 items)

Based on these data, it is estimated that a 40-item vocabulary score and a 36-item reading comprehension score would each have reliabilities exceeding .80 in samples such as those employed in the internal studies cited. Since the validity of a test is partially a function of its reliability, the differences in reliability should be kept in mind in evaluating the validity of the various part scores--that is, a shorter test of a given ability may be expected to have somewhat lower validity than a longer test of that ability, given a common external criterion. For purposes of this study, reliabilities approximating those noted above are assumed to obtain for the various measures.

Preliminary operations on the raw GRE total and part scores. In evaluating the predictive validity of operational GRE verbal, quantitative, and analytical scores, the fact that the scores are based on different forms of the test does not pose problems of score comparability across forms. Through a process of test equating, raw total scores earned on each new form of the GRE General Test are placed on the GRE scale by means of formulas that calibrate the scores to make them comparable with those on earlier forms, regardless of differences in the level of difficulty of the respective forms (for example, ETS, 1981).

However, equating procedures involve only the raw total scores on the respective forms of the test--different sets of item types within a test are not necessarily parallel in difficulty in a given form, and sets of items of a given type are not necessarily parallel in level of difficulty across

*In internal analyses, sentence completions are combined with analogies and antonyms for statistical evaluation.

forms. Thus, combining raw-score data across forms without formal equating introduces some elements of interpretive ambiguity into a validity study. The analysis could have been conducted using only data from a single test form (obviating interpretive complications) but this was not considered desirable because use of single-form data would have substantially restricted sample size and because there might be differences across forms and administrations in examinee mix with respect to variables such as sex, educational status at time of testing, selectivity level of undergraduate school attended, and the like--variables that could have some bearing on study outcomes.

Formal equating of the raw part scores was not feasible for this exploratory study. Without resolving questions regarding the relative difficulty of the respective item types within and/or across forms, it was decided to transform raw part and total scores to a common scale, by form, with full awareness of the attentuation in validity that might be associated with this procedure. In this regard, it was assumed that item types differ only randomly, within and across forms, with respect to parallelism. It was also assumed that attenuating effects due to lack of prallelism were not likely to affect systematically the relative validity of particular sets of items. (See Table 14 and Appendix B for evidence bearing on these assumptions.)

Based on data for examinees taking each form of the GRE General Test without regard to their field of study, raw part and total score distributions were subjected to a z-scale transformation (mean = 0.0, standard deviation = 1.0)--that is, raw part and total scores were expressed as deviations from the respective form grand means in standard deviation units, using the means and standard deviations shown in Appendix A.*

It was reasoned that validity coefficients for the z-scaled part and total scores would be attenuated by any errors associated with lack of equating, while coefficients fdor the GRE scaled (converted, fully equated) total scores would not. It was assumed that comparison of validity coefficients for the total scaled (equated) scores with those for the z-scaled (uequated) total scores would indicate the overall effect on validity of combining unequated total (and part) scores across forms. It was assumed further that, for comparing the validity of total test scores with that of various part scores, the appropriate total scores would be the z-scaled transformations of the raw total scores (paralleling transformations of the respective part scores) rather than the converted GRE scaled total score.

---

*Appendix A also provides data on the number of examinees taking each form by sex. These data indicate pronounced differences in sex mix across forms and administrations; males constituted a majority of examinees taking forms administered in October, December, and February while females constituted a stronger majority of those taking forms administered in April and June. By inference, differences in major-field mix may also be present.

In summary, the test variables available for study following the operations described above were as follows:

V GRE scaled verbal score (equated across forms)
Q GRE scaled quantitative score (as for V)
A GRE scaled analytical score (as for V)
V* Standardized raw total verbal score (not equated by form)
Q* Standardized raw total quantitative score (as for V*)
A* Standardized raw total analytical score (as for V*)

Standardized raw item-type part scores:

ANT (antonyms)
ANA (analogies)
SC (sentence completions)
RD (reading passages)
VO (vocabulary or ANT + ANA)
RC (reading comprehension or SC + RD)

QC (quantitative comparison)
RM (regular mathematics)
DI (data interpretation)

AR (analytical reasoning)
LR (logical reasoning)

Finally, one additional set of GRE "total scores" (designated V#, Q#, and A#, respectively) was included, namely, one in which the various item-type part scores were given equal weight. Given the z-scaled part scores, total verbal, quantitative, and analytical scores defined by the sum of their respective parts were computed for each member of the study sample. In these total scores, item types are equally weighted since the standard deviations of the z-scaled scores are identical. If validity coefficients for V#, for example, should exceed those of, say, V or V* (in both of which the item-type subtests are weighted according to their length), then it may be concluded that the current relative representation of the respective parts in the total score is not consistent with their relative contribution to prediction.

Study Procedures

As indicated earlier, scores on the study variables were available for 437 undergraduate departmental samples, distributed among 12 major fields. In order to assess similarities and differences among the major-field classifications, without regard to department of undergraduate enrollment, profiles of means on the z-scaled item-type part scores were developed for the 12 major-field groups. Questions regarding the relationship of the respective test measures to the GPA-UGPA criterion were explored using scores that were first standardized by department and then pooled across all departments within the respective fields of study.

Pooling rationale. Results of regression analyses in small samples are subject to substantial sampling fluctuation. By pooling data for several small samples from similar settings (for example, several undergraduate chemistry departments), it is possible to obtain more reliable estimates of relationships than would be possible in single small samples. In pooling data across departments one useful approach has been to standardize the predictor and the criterion variables within each department before pooling—that is, to express scores on all variables as deviations from department means in department standard deviation units (see, for example, Wilson, 1979; 1982). For each departmental sample, the mean on each variable is zero and the standard deviation is unity.

Coefficients computed for pooled departmentally standardized variables, by field, may be thought of as approximating population values around which the coefficients for individual departments will vary due to selection- and sampling-related considerations (for example, restriction of range on predictors) as well as context-specific validity-related factors (for example, economics departments may differ in curricular emphasis on quantitative methods of analysis).

A majority of the variation in observed validity coefficients in samples from similar settings tends to be accounted for more by statistical artifacts than by situation-specific validity-related factors. For example, it was found that about 70 percent of the variation across 726 validity studies in the correlation between Law School Admission Test scores and first-year law school grades was attributable to differences in sample standard deviations, estimated criterion reliability, and sample size (Linn, Harnisch, & Dunbar, 1981). Similar findings have been reported for employment settings (for example, Pearlman, Schmidt, & Hunter, 1980).

When analyses are based on pooled, departmentally standardized data within a given field of study, emphasis is on identifying the characteristic patterns of relationships between the respective GRE variables and the measure of academic performance under consideration.

Major-Field Differences in Average Performance on GRE Item-Type Subtests

Table 2 shows means on the GRE verbal, quantitative, and analytical item-type part scores and the respective total scores for examinees in the 12 major fields of study. For all except the converted (GRE scaled) verbal, quantitative, and analytical total score means, the means indicate the average deviation of the raw scores of examinees in a given field from the mean of all examinees in the study sample without regard to field, in all-examinee standard deviation units.

Thus, for example, undergraduate English majors were .622 standard deviations above the all examinee mean on the verbal test (STNRAW V* = .622), .376 standard deviations below the all examinee mean on mean quantitative ability (mean STNRAW Q* = -.376), and so on. Similar interpretations may be made for other means in the table.

Table 2

Means for Major Field Groups on Test Variables

Variable	English	History	Sociology	Political Sci	Chemistry	Computer Sci	Math	Economgr	Economics	Biology	Agriculture	Education
Converted Verbal	590.7	578.2	440.7	536.1	534.3	527.4	546.2	531.2	553.1	537.2	476.3	438.3
Converted Quantitative	533.4	559.1	467.8	549.8	656.4	667.9	700.9	703.6	631.0	605.2	559.1	466.6
Converted Analytical	566.3	576.6	484.6	560.6	611.2	625.9	634.1	616.1	602.0	587.3	543.5	497.4
STNRAW-V*	.672	.502	- 520	.157	.124	.091	.247	.111	.311	.165	-.357	-.672
STNRAW-Q*	-.376	-.187	-.888	-.245	.591	.704	.966	.976	.398	.190	-.155	-.832
STNRAW-A*	-.031	.036	-.695	-.070	.342	.482	.547	.390	.274	.145	-.202	-.546
Antonyms	.671	.543	-.38[illegible]	.172	.041	-.000	.197	.007	.231	.117	-.363	-.546
Analogies	.590	.440	-.459	.166	.096	.092	.228	.040	.249	.112	-.313	-.356
Sentence Completions	.515	.401	-.416	.157	.119	.071	.312	.141	.293	.111	-.336	-.583
Reading Passages	.337	.315	-.505	.064	.167	.142	.139	.191	.284	.206	-.211	-.396
Vocabulary	.699	.546	-.460	.176	.072	.047	.232	.026	.264	.126	-.375	-.605
Reading Comprehension	.451	.385	-.516	.113	.162	.124	.231	.188	.318	.183	-.287	-.650
Quantitative Comparisons	-.320	-.153	-.837	-.217	.562	.652	.924	.926	.369	.173	-.163	-.801
Regular Math	.417	-.249	-.809	-.291	.560	.696	.977	.961	.335	.177	-.164	-.728
Data Interpretation	-.226	-.031	-.663	-.091	.398	.467	.523	.594	.370	.151	-.035	-.663
Analytical Reasoning	-.130	-.041	-.688	-.146	.345	.531	.569	.416	.198	.135	-.139	-.479
Logical Reasoning	.376	.284	-.486	.163	.195	.155	.260	.163	.367	.058	-.286	-.498
N	([illegible])	(584)	(384)	(545)	(644)	(647)	(251)	(850)	(663)	(1319)	(976)	(1649)

Note: Converted V, Q, and A are operational GRE-scaled scores, fully equated across all forms administered during the year. STNRAW V, Q, and A are within-form standardizations of unequated raw total scores on the respective tests. Raw total scores were z-scaled by form using data for all examinees taking each form without regard to their fields of study.

*Except for the converted V, Q, and A scores, all test scores were z-scaled by form of test taken, using data for all examinees taking a form without regard to field. Thus, the grand mean for all examinees is 0.0 and the standard deviation of each test distribution is unity. The means reported indicate the deviation of the mean for a given major field group from the all-examinee mean in standard deviation units. Thus, for example, the mean of .672 for antonyms reported for English majors indicates that they were .672 standard deviations above the grand mean on this variable, on the average.

BEST COPY AVAILABLE

Figure 1 highlights differences among and within fields in performance on the item-type part scores. Profiles for majors in the four humanities and social sciences fields and in education (thought of as verbal fields) are shown together in the left portion of the figure; those for majors in the four math and science fields and economics (thought of as quantitative fields), and in biology and agriculture (thought of as fields with mixed or balanced quantitative and verbal emphases) are shown in the right portion of the figure.

Within-field differences in level of performance on the item-type part scores are of particular interest. For example, majors in the verbal fields typically performed better on the vocabulary items (ANT and ANA) than on the reading comprehension items (SC and RD); they performed better on data interpretation (DI) items than on quantitative comparisons (QC) and regular mathematics (RM) items; and, with the exception of majors in education, they performed at a sharply higher level on logical reasoning (LR) than on analytical reasoning (AR) items.

Majors in chemistry, mathematics, engineering, and computer science tended to exhibit an opposite pattern, with higher performance on reading comprehension items than on vocabulary items, higher performance on quantitative comparisons and regular mathematics than on data interpretation items, and much higher performance on analytical reasoning than on logical reasoning items. Mathematics majors differed from the others in this cluster primarily by performing considerably less well on reading passages (RD) items than on sentence completion (SC) items.

Verbal part-score profiles for majors in economics, biology, and agriculture tended to parallel those for the math and science fields (better on reading comprehension than vocabulary); on the quantitative part scores, their profiles do not exhibit the extreme contrast between quantitative comparisons, regular mathematics, and data interpretation items characteristic of profiles for the math and science majors. With respect to items in the analytical test, agriculture and biology majors, like math and science majors, performed better on analytical than logical reasoning items, but economics majors, like the verbal majors, had a higher logical reasoning than analytical reasoning mean.

Another way of assessing variability in major-field performance on item-type subtests within the respective ability measures is to examine (a) the relative standing of the several major field groups in terms of means on the subtests within a test and (b) the absolute differences in means for various pairs of subtests. For example, for two parallel tests a high degree of consistency in the ranking of field means and relatively small absolute differences in corresponding z-scaled means would be expected; a lower degree of consistency in field ranks combined with larger differences in z-scaled means, on the other hand, would be expected for tests measuring different abilities.

Table 3 shows for pairs of subtests within the respective tests, (a) whether the ranks of the 12 major fields were identical or different and the absolute difference in the ranks when differences were present and (b) the absolute difference in z-score means. The absence of an entry in the rank

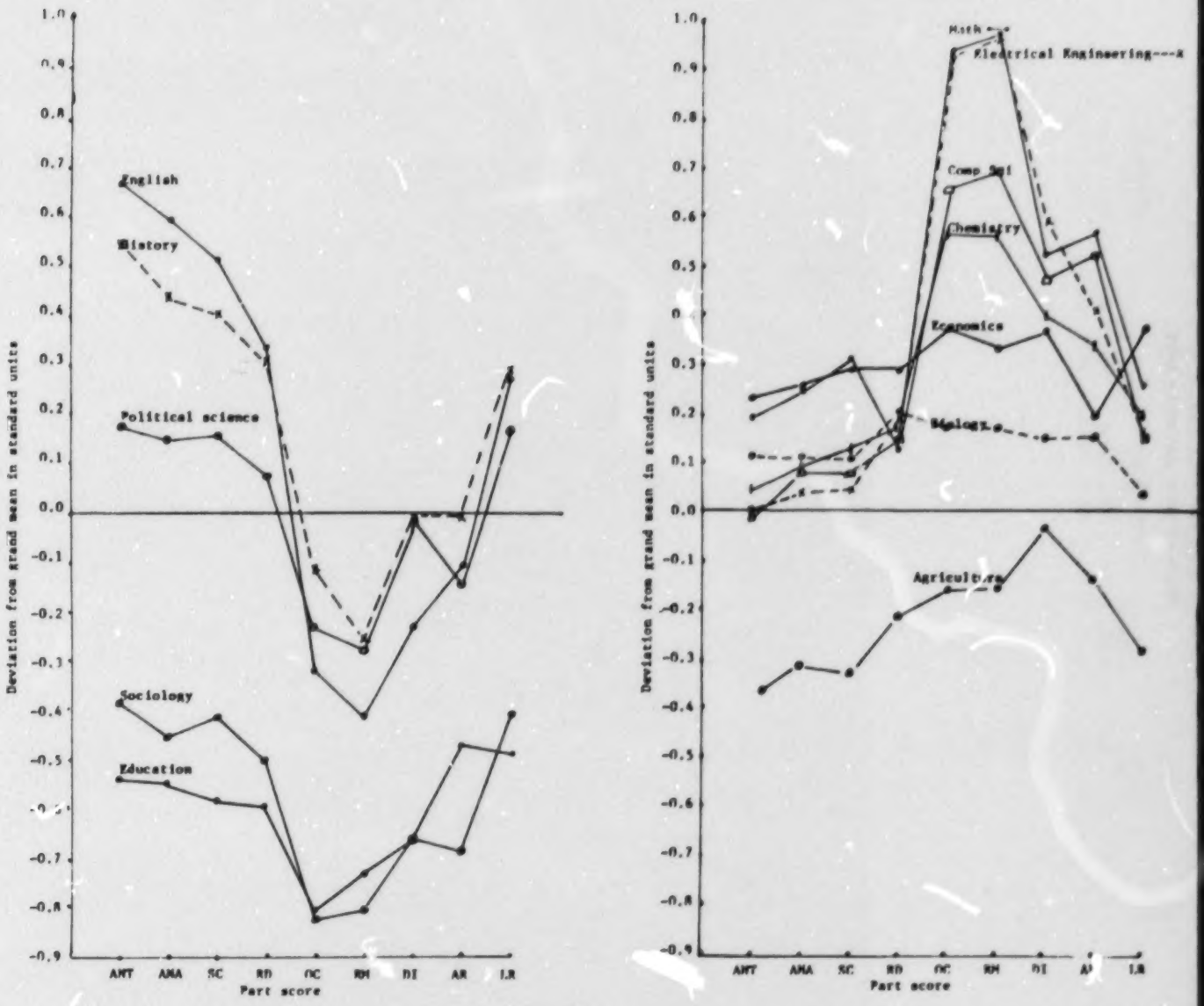


Figure 1. Profiles of mean scores on GRE item-type subtests for undergraduate majors in the fields selected for study

Source: Table 2

Table 3

Observed Absolute Differences in the Ranks of Means of 12 Major Field Groups on Pairs of Item-type Subtests, By Test, and Associated Absolute Differences in Z-score Means

Field	Verbal		Quantitative						Analytical	
	VO-RC		QC-RM		QC-DI		RM-DI		AR-LR	
	Rank Diff.	Mean Diff.	Rank Diff.	Mean Diff.	Rank Diff.	Mean Diff.	Rank Diff.	Mean Diff.	Rank Diff.	Mean Diff.
English		.248	1.0	.090		.092	1.0	.189	5.0	.406
History		.161	1.0	.096		.122	1.0	.218	5.0	.325
Sociology		.056		.028	1.0	.174	0.5	.146	1.0	.292
Pol. Sci.	4.0	.063		.074		.128		.200	3.5	.309
Chemistry		.095		.002		.164		.162	1.0	.150
Comp. Sci.		.077		.044		.185		.229	6.0	.376
Math.		.001	1.0	.053		.401	1.0	.454	3.0	.309
Elec. Eng.	4.0	.162	1.0	.035		.332	1.0	.367	3.5	.253
Economics		.054		.034		.001		.035	4.0	.169
Agriculture		.088	1.0	.001		.128	1.0	.199	1.0	.047
Biology		.057		.034		.022		.026	3.0	.097
Education		.045		.028	1.0	.138	1.0	.064	1.0	.019
Median (diff in mean)		(.067)		(.034)		(.128)		(.175)		(.300)

Note: No entry in the rank difference column indicates that the means for the major field group on the designated pair of subtests had identical rank. If there was any discrepancy in rank, the entry indicates the absolute difference between the ranks. Entries in the mean difference column indicate differences between means in standard units (absolute differences).

Source of data: Table 2

Table 3.1

Intercorrelations of the Ranks of Means on Item-type Part Scores for 12 Major Field Groups

	VO	RC	QC	RM	DI	AR	LR
VO	—	.888	.182	.161	.187	.315	.897
RC		—	.406	.357	.411	.310	.897
QC			—	.986	.999	.958	.403
RM				—	.984	.958	.362
DI					—	.956	.406
AR						—	.481
LR							—

Note: Entries are rank correlation coefficients (rho).

difference column indicates that the ranks of the major field gorup on the designated pair of tests were identical. For additional perspective, Table 3.1 shows rank correlations (rho) of subtest means for the 12 major fields, across as well as within tests.

- o Considering the two verbal subtests, VO and RC, for 10 of the 12 fields, ranks of z-scaled means were identical and the median absolute difference in z-scaled means was relatively small (.067). The rank correlation (rho) for the 12 field means on these two variables was .888 (Table 3.1).

- o For each of the three pairs of quantitative subtests, there were relatively minor discrepancie in rank order, with no shift of more than one rank. For QC and RM, the median absolute difference in means was quite small (.034); however, median differences in field means were greater for both QC and DI (.128) and RM and DI (.175). The average rank correlation of field means (Table 3.1) for the three pairs of quantitative subtests approached .99.

- o For the two analytical ability subtests, AR and LR, some shifts in ranking were found for every field, the median absolute difference in z-scaled means (.300) was higher than that for subtests within the verbal and quantitative ability measures, and the rank correlation of field means on AR and LR (rho = .481) was considerably lower than that for the other pairs of subtests.

The findings regarding field means indicate the differential development within individuals, associated with field of concentration, of the skills and abilities being measured by different item types within the respective tests. On balance, the evidence reviewed in this section suggests that the item-type part scores are not simply different methods of measuring their respective constructs but that they may represent distinguishable components of underlying general abilities with the potential for independent measurement utility.

In this connection it is important to note (a) that the degree of consistency in major field performance differentials is greater for subtests within the verbal and quantitative ability measures than for the two analytical ability subtests, (b) that the field ranks on analytical reasoning items correspond closely with ranks on all three quantitative item types (average rho of approximately .960), and (c) that field ranks on logical reasoning items correspond closely with ranks on the two verbal subtests (rho = .8967 for both LR-VO and LR-RC). Generally speaking, the rank correlations in Table 3.1 indicate that, insofar as major field performance differences are concerned, the information conveyed by the analytical reasoning and the quantitative subtests is similar and that conveyed by the logical reasoning subtest and verbal subtests is also similar.

Exploratory Evaluation of Part-Score Validity

The analyses involving part-scores on the verbal measure were guided by

several a priori working hypotheses, based on the College Board findings cited at the outset:

1. The GRE reading comprehension (RC) subtest (based on sentence completions and reading comprehension sets) should be more closely related to SR-UGPA than the GRE vocabulary (VO) subtest (based on antonyms and analogies).

2. The 36-item RC subtest should be comparable in validity to the total GRE verbal test, including the 40 VO items.

3. The multiple correlation of the RC,Q*,A* battery with SR-UGPA should be comparable to that of the V*,Q*,A* battery.

4. Occasional suppression of VO, but not [illegible]ance may be expected in composites including RC, VO, and other GRE va[illegible]s.

In the absence of comparable working hypotheses regarding the quantitative and analytical part scores, evaluation of observed relationships for these item types was guided by interest in (a) the relative contribution of the respective item-type part scores within each test to prediction of SR-UGPA, (b) the comparative validity of total test scores and the component part scores, and (c) evidence suggesting the possibility that separately scored item-type subtests might provide a basis for improved assessment.

The Verbal Test Part-Score Analysis

Table 4 shows pooled within-department correlations between SR-UGPA and (a) VO and RC scores, (b) various verbal total scores, and (c) a best-weighted combination of VO and RC scores, by field, and for all fields combined. Validity coefficients for V* (the raw unequated total verbal score, z-scaled by test form) were slightly lower than those for V (the converted, GRE-scaled operational verbal score). This outcome is expected because V* total scores, like the respective part scores, were not equated across forms. In comparing part- and total-score validity, V* is judged to be the more appropriate total, under the assumption that attenuating effects associated with lack of equating across test forms are comparable fdor V* and the respective part scores. Coefficients for V# (a total defined as the sum of equally weighted scores on analogies, antonyms, sentence completions, and reading sets) and V[illegible] are assumed to be comparably attenuated due to errors associated with lack of equating across forms. This same line of reasoning is applicable, of course, to later consideration of data on the quantitative and analytical measures.

The validity coefficients for the verbal measure varied by field generally in accordance with the expectation of higher validity in the more verbal fields than in the more quantitative fields. This was true without regard to the particular verbal measure under consideration. However, for

Table 4

Pooled Within-Department Correlations of Selected Verbal Part and Total Scores with SR-UGPA, by Field

Field	(N)	Verbal part scores		Verbal total scores				Difference in validity		
		VO r	RC r	V* r	V# r	V r	VO,RC R	RC vs VO	V* vs RC	VO,RC vs V*
English	**(884)**	**347**	**377**	**395**	**395**	**399**	**399**	**030**	**018**	**004**
History	**(584)**	**322**	**354**	**366**	**370**	**377**	**375**	**032**	**012**	**009**
Sociology	**(364)**	**342**	**396**	**407**	**384**	**418**	**417**	**054**	**011**	**010**
Polit Sci	**(545)**	**283**	**376**	**362**	**364**	**364**	**380**	**093**	**-014**	**018**
Chemistry	**(644)**	**226**	**217**	**243**	**242**	**249**	**248**	**-016**	**026**	**005**
Computer Sci	**(647)**	**238**	**213**	**246**	**209**	**245**	**251**	**-025**	**033**	**005**
Mathematics	**(251)**	**248**	**312**	**292**	**296**	**321**	**314**	**064**	**-020**	**022**
Elec Engin	**(850)**	**140**	**249**	**211**	**201**	**223**	**253**	**109**	**-038**	**042**
Economics	**(663)**	**323**	**391**	**391**	**390**	**404**	**403**	**068**	**000**	**012**
Biology	**(1318)**	**228**	**288**	**286**	**269**	**302**	**297**	**060**	**-002**	**011**
Agriculture	**(976)**	**214**	**215**	**239**	**236**	**260**	**239**	**001**	**024**	**000**
Education	(1649)	296	313	333	326	332	335	007	010	002
All Fields	(9375)	263	301	309	300	318	315	038	008	006

Note. V* is the raw total verbal score, z-scaled by form.

V# is an equally weighted sum of four verbal part scores.

V is the converted GRE scaled verbal score, equated across forms.

VO,RC is a best weighted composite of the designated part scores.

Entries are correlation coefficients without decimals.

economics, among the more quantitative fields, the verbal test had validity coefficients comparable to the coefficients for the English, history, sociology, and political science samples.

With respect to patterns of verbal part-score validity, the findings in Table 4 are generally consistent with the basic working hypotheses outlined above.

- o Considering first the all-fields coefficients (equivalent to weighted averages of coefficients for 437 departments without regard to field), the validity for RC is greater than that for VO (coefficients differ by .038, as indicated in the RC vs VO difference column), and this was true for 10 of the 12 fields. For chemistry and computer science departments, the mean VO coefficient was higher than the mean RC coefficient, but the mean difference was less than the average for all departments in absolute magnitude.

- o RC alone was about as valid as V* including the VO items. Considering data for all departments, without regard to field, coefficients were .301 and .309, respectively. RC was actually slightly more valid than V* in several fields.

- o However, a best-weighted composite of VO and RC did not yield much better prediction than the V* total score, similar to the results observed with SAT vocabulary and reading comprehension when they were similarly treated. Largest differences in validity between the VO,RC composite and V* occurred in three of the four fields in which RC was more valid than V*, and in which differences in validity between RC and VO were greatest, namely, electrical engineering, mathematics, and political science.

The data in Table 5 lend support to the working hypothesis that the multiple correlation of an RC,Q*,A* composite with SR-UGPA should be comparable to that of a V*,Q*,A* composite.

- o For all departments, without regard to field, the coefficient for RC,Q*,A* was only .002 points less than that for V*,Q*,A*, and .005 points less than that for VO,RC,Q*,A*--that is, when VO was added to the RC,Q*,A* battery there was little increase in the multiple correlation. There were no notable exceptions to this general finding by field.

Table 6 provides evidence regarding the relative weighting of two sets of verbal part scores, namely, VO and RC (Set 1), and the four basic verbal item types, namely, analogies, antonyms, sentence completions, and reading comprehension sets (Set 2), when included in a battery with Q* and A*.

- o The data in Set 2 indicate, among other things, (a) that, over all departments, the relative weighting of sentence completions and reading items (components of the RC score) was approximately equal, (b) that one of these two RC item-types was the highest of the four verbal item types in all fields but one (agriculture), but (c) that the relative weighting of the SC and RD items, when they were allowed to compete independently, varied across fields without regard to their verbal or quantitative emphasis.

Table 5

Multiple Correlation with UGPA of Quantitative, Analytical, and Selected Verbal Scores, by Field

		Multiple correlation (1) RC,Q*,A*	(2) VO,RC Q*,A*	(3) V*,Q*,A*	Difference (3-1)	(3-2)	(2-1)
English	(884)	384	403	402	018	-001	019
History	(584)	362[a]	382[a]	374[a]	012	-008	020
Sociology	(364)	436	447	442	006	-005	011
Political Sci	(545)	419[a]	420[a]	410[a]	-009	-010	001
Chemistry	(644)	372	375	374	002	-001	003
Computer Sci	(647)	365	374	371	006	-003	009
Mathematics	(251)	412[a]	412[ab]	395[a]	-017	-017	000
Elec Engin	(850)	393	406[a]	386	-007	-020	013
Economics	(663)	452	458	455	003	-003	006
Biology	(1318)	352	354	350	-002	-004	002
Agriculture	(976)	299	306	306	007	000	007
Education	(1649)	356	366	366	010	-001	001
All Fields	(9375)	361	366	363	002	-003	005

Note: Entries are multiple correlation coefficients or differences between designated coefficients without decimals.

VO = ANT + ANA = Vocabulary
RC = SC + RD = Reading Comprehension

V*, Q* and A* are raw total scores on the respective tests, z-scaled by form.

[a] A* variance is suppressed in this composite.

[b] VO variance is suppressed in this composite.

Table 6

Relative Weighting of Two Sets of Verbal Part Scores, Quantitative and Analytical Scores in Composites for Predicting UGPA, by Field

Field	(N)	Set 1 Beta weights				(R)	Set 2 Beta weights						(R)
		VO	RC	Q*	A*		ANA	ANT	SC	RD	Q*	A*	
English	(884)	166	230	061	017	(403)	104	077	168	168	063	010	(411)
History	(584)	159	238	089	-045	(382)	037	101	241	072	078	-039	(398)
Sociol	(364)	126	215	035	171	(447)	142	031	056	168	034	169	(451)
Pol Sci	(545)	038	256	232	-054	(420)	093	-045	108	180	230	-060	(425)
Chem	(644)	064	031	279	072	(375)	020	054	007	030	277	072	(376)
CS	(647)	106	011	274	064	(374)	059	063	-064	065	274	059	(376)
Math	(251)	-012	237	300	-034	(412)	-128	077	209	098	296	-016	(431)
Elec E	(850)	-147	179	320	065	(406)	-089	-061	068	123	320	066	(406)
Econ	(663)	099	196	144	143	(458)	040	066	102	129	143	145	(458)
Biol	(1318)	050	158	187	054	(354)	028	036	033	142	189	052	(357)
Agric	(976)	085	038	178	076	(306)	026	065	050	009	179	073	(309)
Educ	(1649)	122	131	148	048	(366)	099	040	103	053	142	048	(371)
All Fields	(9375)	079	145	185	047	(366)	044	044	079	090	185	047	(366)

Note. Entries are standard partial regression (beta) weights and multiple correlation coefficients without decimals.

VO = ANT + ANA = Vocabulary; RC = SC + RD = Reading Comprehension

Q* and A* are raw total test scores, z-scaled by form.

Negative weights reflect suppression of variance; zero-order coefficients are positive.

Underscored weights are estimated to be significant at the .05 level.

- o Suppressor effects, indicated by negative regression weights for predictors that are positively correlated with a criterion, were present for VO and/or VO component item types in analyses for mathematics, engineering, and political science departments, consistent with the hypothesis of occasional suppressor effects for vocabulary items; in one analysis (computer science departments), the sentence completion subtest was negatively weighted, contrary to hypothesis.

- o The Set 1 and Set 2 multiple correlations are identical in the analysis over all departments and are almost so in the respective field analyses.

The analyses reviewed above indicate differences in the criterion-related validity of the VO and RC subtests favoring the RC subtest, which appears to be carrying most of the predictive validity load in the total verbal score when the criterion is SR-UGPA.

The Quantitative Test Part-Score Analysis

Table 7 provides data on the relationship of the three quantitative item-type part scores to SR-UGPA. The correlations of three quantitative total scores, namely, Q*, Q#, and Q, with the same criterion are also shown. As expected, the validity coefficients for the various quantitative total scores are higher for the math and science and economics departments than for the other, less quantitative fields; however the higher validity of quantitative scores for political science departments than for other verbal departments was not expected.

In the absence of an a priori basis for expecting particular patterns of differential validity for the respective item types, perhaps the most relevant general consideration to be kept in mind is that the three quantitative subtests differ in length. QC includes 30 quantitative comparison items, RM includes 20 regular mathematics items, and DI includes 10 data interpretation items. Thus, we would expect validity coefficients to vary with test length if the three item-types are actually homogeneous with respect to the abilities they tap.

- o For all departments, the validity patterns for QC, RM, and DI followed the variation-according-to-length hypothesis, and this was true for several of the field analyses as well. However, there were exceptions. For example, RM validities were somewhat higher than those for QC in several fields, most notably so in mathematics; DI validities comparable to those for QC were obtained in analyses for agriculture, English, and sociology (which are among the fields in which students performed better on the DI subtest than on the QC and RM subtests--see Figure 1).

- o A composite of the separately weighted part scores did not result in better prediction than that provided by Q*, based on the analysis over all departments. Only in the analysis for mathematics departments, in which the regular mathematics items had uniquely high validity, was there a notable exception to the foregoing generalization. The content

Table 7

Pooled Within-Department Correlations of Quantitative Part and Various Total Scores with UGPA, by Field

Field	(N)	Quantitative part scores			Quantitative total scores			
		QC	RM	DI	Q*	Q#	Q	QC,RM, DI
		r	r	r	r	r	r	R
English	(884)	209	209	192	238	241	246	245
History	(584)	186	203	126	212	205	225	224
Sociology	(364)	226	259	216	285	293	310	286
Polit Sci	(545)	353	2.9	216	353	335	362	362
Chemistry	(644)	330	305	203	358	340	371	366
Computer Sci	(647)	285	293	258	350	349	356	350
Mathematics	(251)	294	366	210	356	340	378	382
Elec Engin	(850)	346	290	212	378	356	397	380
Economics	(663)	307	283	212	348	339	358	348
Biology	(1318)	268	246	182	296	287	310	298
Agriculture	(976)	216	204	217	276	280	306	278
Educ	(1649)	285	243	195	302	292	302	304
All Fields	(9375)	274	257	201	308	300	320	308

Note. Q* is the raw total quantitative score, z-scaled by form.
Q# is an equally weighted sum of quantitative part scores.
Q is the converted quantitative score, equated across forms.
Entries are correlation coefficients without decimals.

of the regular mathematics items may overlap more with the content of the major field for mathematics majors than for majors in the other fields. If so, this would help to explain the strong predictive validity of these items and would be consistent with findings of previous research indicating characteristically higher validity for the GRE Subject (Advanced) Tests than for the General (Aptitude) Test (see, for example, Willingham, 1974; Wilson, 1979; 1982).

Table 8 provides insight into the relative weighting of QC, RM, and DI when the three part scores were included in a battery with A* and V*. The difference in multiple correlation between the QC,RM,DI,A*,V* composite and the Q*,A*,V* composite is also shown. The predictive load, relative to the SR-UGPA criterion, in the quantitative test is being borne primarily by the QC and RM items, judging from the findings in Table 8.

- o DI contributed only slightly to prediction, generally, and attained statistical significance only in the analyses for computer science and agriculture departments; suppression effects were found for DI in analyses for two verbal fields (history and political science) and mathematics. In a stepwise regression program, QC, RM, and DI were entered as a set followed sequentially by the introduction of A*, then V*. In the three analyses showing DI suppression (and in all other analyses), the weight for DI was positive in the initial quantitative set. The DI weight became negative only after the introduction of the final variable (V*) in analyses for history and political science, but after the introduction of A* in the mathematics analysis.
- o Separate treatment of QC, RM, and DI part scores in a battery with A* and V* did not lead to better prediction than that provided by Q*,A*,V* (see difference column in Table 8).

The Analytical Test Part-Score Analysis

The analytical ability measure introduced in October 1981 is a revised version of the analytical measure introduced when the GRE General Test was restructured in 1977. There is empirical evidence regarding the validity of the October 1977 analytical measure for predicting graduate school performance (for example, Wilson, 1982), but evidence regarding the October 1981 version is more limited. Evidence of positive relationships between SR-UGPA and analytical reasoning and logical reasoning items, respectively, was reported by Wild, Swinton, and Wallmark (1982) in studies leading to the revision of the 1977 measure. In those studies, logical reasoning items were found to be more closely related to SR-UGPA than analytical reasoning items in samples that were not differentiated with respect to field.

The analyses reported in this section provide evidence regarding the relationship of the various analytical ability total scores (A*, A#, and A) and the component analytical ability item types, namely, analytical reasoning (AR) and logical reasoning (LR), to SR-UGPA in samples classified

Table 8

Relative Weighting of Quantitative Item-Type Part Scores in a Composite with A* and V*

Field	(N)	Beta weights					QC,RM,DI A*,V* (R)	Increase over Q*,A*,V*
		QC	RM	DI	A*	V*		
English	(884)	-004	045	054	018	355	404	003
History	(584)	053	096	-033	-040	347	380	006
Sociology	(364)	-037	058	013	182	301	444	002
Polit Sci	(545)	223	050	-007	-041	253	416	006
Chemistry	(644)	184	132	028	074	077	381	007
Computer Sci	(647)	116	126	100	061	101	370	000
Mathematics	(251)	094	265	-004	-021	191	418	023
Elec Engin	(850)	219	124	047	082	030	389	003
Economics	(663)	091	068	002	159	260	454	000
Biology	(1318)	126	089	008	063	180	352	002
Agriculture	(976)	042	095	087	073	114	308	002
Education	(1649)	117	049	002	052	226	368	002
All fields	(9375)	107	089	025	053	197	364	001

Note. Entries are standard partial regression (beta) weights or multiple correlation coefficients without decimals.

Underscored weights are estimated to be statistically significant ($p < .05$).

by field of study. In evaluating the observed correlations, in Table 9, it is important to keep in mind that total scores on the 50-item analytical measure are more heavily influenced by performance on the 38 analytical reasoning items than by performance on the 12 logical reasoning items.

Generally speaking, typical validity coefficients for the various analytical total scores tend to be somewhat higher in the primarily quantitative fields (except mathematics) than in the verbal fields (except sociology). However, the AR and LR subtests do not have similar patterns of validity coefficients across verbal and quantitative fields.

In this regard, perhaps the most striking aspect of the part-score validity data in Table 9 is the strong contribution to prediction of GR-UGPA, relative to that of the 38-item AR subtest, of the LR subtest based on only 12 logical reasoning items.

- For all departments, the validity of the LR subtest was .225 as compared to .229 for the longer AR subtest.
- In seven analyses, the validity coefficient for LR was approximately equal to or greater than that for AR.
- In three analyses (for history, political science, and education departments), the LR subtest validity coefficient was greater than that for the A* total (which included the AR items).
- AR validities tended to be somewhat higher for the basically quantitative fields than for the basically verbal fields; for LR, validity coefficients tended to show an opposite pattern.

The relative weighting of AR and LR in an independently computed composite and their weighting in a composite with V* and Q* are shown in Table 10.

- When AR and LR were treated as predictors, AR weights were somewhat higher than those for LR in composites for the chemistry, computer science, mathematics, electrical engineering, biology, and agriculture analyses.
- LR weights were somewhat higher than AR weights in analyses for history and political science (among the more verbal fields), for economics alone among the more quantitative fields, and for education.

Although, when considered jointly as an independent battery, weights for both AR and LR reached the .05 level of statistical significance in most of the analyses, neither AR nor LR made a consistent, substantial contribution to prediction when treated as elements in a battery that included the V* and Q* total scores (cf. results in Table 6 for verbal subtests combined with A* and Q* and in Table 8 for quantitative subtests combined with V* and A*).

- Only the beta weight for LR was significant in the overall departmental analysis, and its contribution to prediction was relatively slight (beta = .058 as compared to approximately .190 and .185 for V* and Q*).

Table 9

Pooled Within-Department Correlations of Analytical Part Scores and Various Total Scores with UGPA, By Field

Field	(N)	Analytical part scores		Analytical total scores			
		AR	LR	A*	A#	A	AR,LR
		r	r	r	r	r	R
English	(884)	202	200	236	248	239	243
History	(584)	146	239	195	233	205	249
Sociology	(364)	314	300	360	376	375	372
Polit Sci	(545)	162	269	229	272	232	278
Chemistry	(644)	255	175	275	262	282	270
Computer Sci	(647)	224	221	259	267	256	266
Mathematics	(251)	218	214	239	250	251	263
Elec Engin	(850)	264	199	282	273	292	287
Economics	(663)	301	335	358	386	361	388
Biology	(1318)	224	179	244	242	251	250
Agriculture	(976)	220	184	240	238	255	246
Education	(1649)	234	256	242	296	279	297
All Fields	(9375)	229	225	264	274	270	275

Note. A* is the raw total analytical score, z-scaled by form.
A# is an equally weighted sum of the analytical part scores.
A is the converted analytical score, equated across forms.
AR,LR is a best weighted composite of the designated part scores.
Entries are correlation coefficients without decimals.

Table 10

Relative Contribution of AR and LR to Prediction of SR-UGPA in an Independent Composite and in a Composite Including Q* and A*

Field	(N)	Beta weights		Multiple correlation	Beta weights				Multiple correlation
		AR	LR	(R)	AR	LR	Q*	V*	(R)
English	(884)	148	145	(243)	012	013	070	353	(401)
History	(584)	074	214	(249)	-063	071	097	319	(381)
Sociology	(364)	241	214	(372)	133	107	031	284	(444)
Polit Sci	(545)	078	242	(278)	-102	088	248	231	(422)
Chemistry	(644)	221	094	(270)	069	-003	283	089	(374)
Computer Sci	(647)	160	157	(266)	-002	092	289	083	(376)
Mathematics	(251)	163	158	(263)	-010	024	288	175	(395)
Elec Engin	(850)	221	119	(287)	065	042	314	024	(387)
Economics	(663)	209	261	(388)	094	127	147	230	(461)
Biology	(1318)	185	116	(250)	044	039	191	173	(351)
Agriculture	(976)	176	119	(246)	054	047	178	103	(307)
Education	(1649)	161	197	(297)	008	080	154	206	(370)
All Fields	(9375)	170	164	(275)	022	058	190	185	(365)

Note. Decimal points have been omitted from all coefficients. Underscoring indicates estimated statistical significance at the $p < .05$ level.

- o Suppression effects were found for AR in four departmental analyses and for LR in one; in the samples involved, AR or LR criterion-related variance was more than sufficiently represented in the verbal and/or quantitative total scores.

- o Weights for both AR and LR were statistically significant in only two field analyses (sociology and economics) and LR was significant in a third (education).

The data in Table 10 suggest that the analytical test, as currently defined by the 38 AR and 12 LR items, is not providing very much unique, SR-UGPA-related information. This conclusion is reinforced by the data in Table 11, which permit comparison of multiple correlations with SR-UGPA of V*Q* only and those yielded by adding A* and AR and LR, respectively. Increments in R due to adding analytical test scores to V* and Q* typically were quite small. In evaluating this finding, it is useful to know that V*Q* alone yielded a higher multiple correlation with SR-UGPA than either A*V* or A*Q* in 9 of the 12 field analyses and in the total sample.

Understanding of these findings is advanced by reference to Table 12 and Table 12.1. In Table 12 it may be seen that LR is more closely related to a verbal subtest (RC) than to LR. From Table 12.1 it may be determined that the average within-test intercorrelations of verbal subtests (.503) and quantitative subtests (.476) are greater than that observed for the two analytical ability subtests (.360); moreover, the correlation of LR with three of the four verbal subtests is higher than that of AR with these subtests while the correlation of AR with each quantitative subtest is higher than that of LR with these subtests. Intercorrelations corrected for errors of measurement shown in Table 12.1 (below the diagonal) lead to similar conclusions. In essence, AR items tend to have more in common with quantitative items than with LR items, while LR items have more common variance with verbal items than with AR items.

Verbal, Quantitative, and Analytical Part Scores as a Battery

Table 13 shows major findings of an analysis of the regression of SR-UGPA on seven item-type part scores, namely, VO, RC, QC, RM, DI, AR, and LR. Standard partial regression (beta) weights are shown for variables selected by stepwise regression as contributing at least .001 to R-squared.

- o The consistent significant contribution to prediction of the RC and/or VO subtests is noteworthy; both are significant in four analyses, RC only is significant in five, and VO only in three (though acting as a suppressor in one).

- o The part score that appears to be contributing least to the battery is data interpretation (DI). However, the score for this subtest met the statistical significance criterion in the computer science and agriculture analyses.

Table 11

Incremental Contribution of the Analytical Measure (A*) in Part-Score and Total-Score Form to Prediction of 5k-UGPA After Taking V* and Q* into Account, by Field

Field	(N)	Composite predictor (1) V*,Q* (R)	(2) V*,Q*, A* (R)	(3) V*,Q* AR,LR (R)	Difference in R (2-1)	(3-1)
English	(884)	401	401	401	000	000
History	(584)	374	374[c]	381[a]	000	007
Sociology	(364)	420	442	444	020	024
Polit Sci	(545)	408	410[c]	422[a]	002	014
Chemistry	(644)	370	374	374[b]	004	004
Computer Sci	(647)	368	371	376[a]	003	008
Mathematics	(251)	395	395[c]	395[a]	000	000
Elec Engin	(850)	381	386	387	005	006
Economics	(663)	439	455	461	016	022
Biology	(1318)	346	350	351	004	005
Agriculture	(976)	300	306	307	006	007
Education	(1649)	364	366	370	002	006
All Fields	(9375)	361	363	365	002	004

Note. Entries are correlation coefficients without decimals.

[a] AR negatively weighted

[b] LR negatively weighted

[c] A* negatively weighted

Table 12

Intercorrelations of Analytical Test Part Scores, and their Correlations with Selected Verbal and Quantitative Test Part Scores, by Field

Field	(N)	AR score vs LR score	AR score vs RC	AR score vs QC	LR score vs RC	LR score vs QC
		r	r	r	r	r
English	(884)	373	444	<u>501</u>	<u>449</u>	326
History	(584)	337	404	<u>497</u>	<u>472</u>	321
Sociology	(364)	341	401	<u>502</u>	<u>442</u>	276
Polit Sci	(545)	352	436	<u>476</u>	<u>468</u>	404
Chemistry	(644)	369	415	<u>436</u>	<u>467</u>	295
Computer Sci	(647)	404	416	<u>422</u>	<u>483</u>	226
Mathematics	(251)	346	382	<u>437</u>	<u>512</u>	290
Elec Engin	(850)	363	407	<u>451</u>	<u>484</u>	346
Economics	(663)	358	387	<u>461</u>	<u>508</u>	366
Biology	(1318)	336	<u>409</u>	408	<u>384</u>	238
Agriculture	(976)	368	451	<u>478</u>	<u>488</u>	321
Education	(1649)	366	469	<u>557</u>	<u>507</u>	372
All Fields	(9375)	360	429	<u>475</u>	<u>469</u>	318

Note: Entries are correlation coefficients without decimals.
The higher coefficient in a given comparison is underscored.

Table 12.1

Pooled Within-Department Intercorrelations of Item-Type Part Scores: Total Sample

	ANT	ANA	SC	RD	QC	RM	DI	AR	LR
ANT	--	528	493	486	277	266	233	282	356
ANA	843		509	486	337	290	260	335	371
SC	711	784	--	519	336	274	267	332	394
RD	608	650	749	--	360	322	316	408	426
QC	433	372	485	450	--	548	440	475	318
RM	343	400	378	415	707	--	440	487	310
DI	336	359	388	456	635	651	--	415	264
AR	441	441	407	510	594	628	600	--	360
LR	514	549	593	615	459	462	440	519	--

Note: Values above the diagonal are observed correlations; those below are corrected for errors of measurement by use of the formula $r_{ab}/\sqrt{r_{aa}r_{bb}}$; reliabilities are estimated roughly.

Entries are correlation coefficients without decimals.

Table 13

Beta Weights for Subsets of Item-Type Part Scores Selected by Stepwise Regression According to a Contribution to R^2 Criterion, By Field

Field	Part-score beta weights VO	RC	QC	RM	DI	AR	LR	Selected Set (R)	V*,Q*,A* (R)	N
English	17	23		04	05			406	402[a]	(884)
History	15	22	04	09		-08	06	392	374[a]	(584)
Sociology	12	21		06		11	10	451	442[ac]	(364)
Political Science		26	22	04		-10	08	437	410[ab]	(545)
ALL VERBAL	12	23	05	06			05	(404)	(396[ab])	(2377)
Chemistry	08		19	14		08		381	374[b]	(644)
Computer Science	09		13	13	10		09	377	371[ab]	(647)
Mathematics		22	08	26				434	395[ab]	(251)
Electrical Engin	-14	19	22	13	04	06		409	386[b]	(850)
Economics	09	18	08	07		09	12	463	455[abc]	(663)
ALL QUANTITATIVE		12	15	13	04	05	06	(391)	387[abc]	(3055)
Biology	05	17	13	09		05		356	350[ab]	(1318)
Agriculture	10		05	10	09	06	05	309	306[ab]	(976)
ALL BALANCED	06	11	09	09	04	05	04	(332)	(330[abc])	(2294)
Education	11	12	12	06			08	373	366[ab]	(1649)
ALL FIELDS	07	14	12	10			06	(367)	(363[abc])	(9375)

Note. Entries are regression and correlation coefficients without decimals. The regression coefficients tabled are for part scores contributing at least .001 to R-squared; underscored coefficients also met a .05 statistical significance criterion.

[a]V* significant, .05; [b]Q* significant, .05; [c]A* significant, .05

- o The regular mathematics subscore contributed at least .001 to R-squared in every analysis and is the only part score for which this was true.

- o AR and/or LR were selected as part of the most efficient part-score battery in 10 of the 12 field analyses (though with AR variance suppressed in two).

Generally speaking, the best weighted composites of selected part scores yielded somewhat higher multiple correlations with SR-UGPA than the three total test scores; no corrections for shrinkage have been made, however. In evaluating the findings in TAble 13, it is important to note that the subtests involved are of differing lengths and reliabilities, that the analysis did not attempt tc adjust for these factors, and that, given moderately intercorrelated predictors such as those involved in the analysis, regression weights are sensitive to relatively small changes in validity.

Comparability of Regression Results for Unequated and Equated Total Scores

The preceding analyses were based primarily on test scores that were not equated across test forms. To what extent do patterns of findings based on unequated score data provide a basis for projecting results that might be obtained if equated part and total scores were to be employed? Table 14 presents findings bearing on the comparability of regression results for unequated (V*,Q*,A*) and equated (V,Q,A) total scores on the respective tests.

While there are differences in detail in the results of the parallel analyses, the relative weighting of the verbal, quantitative, and analytical scores, and the relative magnitudes of the multiple correlation coefficients, by field, are essentially the same for the two analyses. It seems reasonable to infer that comparable results might be expected for parallel analyses involving equated and unequated part scores (see Appendix B).

From Table 14 it may be determined that the multiple correlations for the V*,Q*,A* composites are somewhat lower than those for the V,Q,A composites due, it is assumed, to error associated with lack of equating for V*, Q*, and A* across forms.

Summary of Trends in Findings

Major trends in the findings bearing on the predictive and/or construct validity of item-type part scores ar°e summarized below, by test.

With respect to the verbal ability measure—

- o Althought there are some exceptions, by field, reading comprehension items (SC + RD) tend to be more valid than vocabulary items (ANT + ANA) and the same tends to be true of the RC and VO component item types.

- o RC and VO item types appear to be contributing to the prediction of

Table 14

Comparisons of Regression Results for Unequated Raw Total Scores (V*, Q*, A*) and GRE Scaled Scores (V, Q, A), by Field

		Unequated Scores Beta Weights				Equated Scores Beta Weights			
	(N)	V*	Q*	A*	(R)	V	Q	A	(R)
English	(884)	353	066	026	(402)	354	075	025	(407)
History	(584)	345	097	-035	(374)	353	115	-041	(388)
Sociology	(364)	296	025	189	(442)	294	061	184	(459)
Polit Sci	(545)	256	242	-044	(410)	253	256	-047	(417)
Chemistry	(644)	080	281	073	(374)	085	294	074	(389)
Computer Sci	(647)	103	276	060	(371)	108	287	051	(377)
Mathematics	(251)	193	296	-024	(395)	217	308	-022	(425)
Elec Engin	(850)	030	317	081	(386)	040	336	075	(405)
Economics	(663)	258	147	153	(455)	272	159	143	(468)
Biology	(1318)	178	191	062	(350)	195	204	058	(370)
Agriculture	(976)	112	177	074	(306)	125	206	066	(335)
Education	(1649)	226	149	050	(366)	224	149	056	(367)
All Fields	(9375)	196	187	052	(363)	204	201	049	(377)

Note: Entries are standard partial regression (beta) weights or multiple correlation coefficients without decimals.

V*, Q*, A* are raw unequated total scores, z-scaled by form.
V, Q, A are GRE scaled scores, fully equated across forms.

academic performance in fields that vary widely in apparent verbal emphasis.

- o Majors in verbal fields tend to perform better on VO than on RC while majors in quantitative fields tend to perform better on RC than VO (with the anomalous exception of mathematics (see Figure 1 and related discussion).

With respect to the quantitative ability measure--

- o Data interpretation (DI) items appear to be contributing only slightly to overall predictive validity.
- o Regular mathematics (RM) items may be particularly predictive of performance in mathematics (hypothetically, because of a greater degree of overlap between test content and curricular content for mathematics majors than for others).
- o Both RM and quantitative comparisons (QC) items appear to be contributing to prediction, though not necessarily equally so, in fields that differ widely in apparent quantitative emphasis.
- o Majors in verbal fields (for example, history, English, political science) tend to perform much better on DI items than on other quantitative item types, while the opposite is true for majors in math and science fields (for example, engineering, chemistry, computer science, mathematics).

With respect to the analytical ability measure--

- o Based on their relative contribution to prediction of SR-UGPA, logical reasoning (LR) items appear to be underrepresented and analytical reasoning (AR) items overrepresented in the current 12-item to 38-item, LR to AR, ratio in the analytical ability measure. The shorter LR subtest appears to be as valid as the longer AR subtest.
- o Analytical reasoning items behave more like quantitative ability items while logical reasoning items behave more like verbal ability items--they may prove to be useful extensions of the two basic ability measures.
- o Majors in verbal fields perform better on LR than on AR, while the opposite is true for majors in quantitative fields; ranks of fields in terms of mean total analytical ability score differ considerably from ranks based on AR and LR means, and ranks based on AR means differ from ranks based on LR means.

Discussion

Findings regarding the GRE vocabulary and reading comprehension subtests tend to confirm and extend findings based on parallel subtests of

the SAT verbal measure. These results, combined with results of factor studies indicating distinguishable "vocabulary" and "reading comprehension" verbal factors defined by items like those in the subtests under consideration in this study, suggest a potentially useful role for VO and RC subscores as defined for the study.

No a priori rationale was available for projecting particular patterns of validity for item-type part scores on the quantitative and analytical ability measures. However, results suggested that the respective part scores are measuring somewhat different aspects of quantitative and analytical reasoning ability. Based on observed patterns of validity coefficients for quantitative subtests and average scores for different majors, the components of quantitative ability being measured by the data interpretation items appear to be different from those being measured by QC and RM. This is consistent with the results of a factor analysis of small sets of items from the 1977 GRE Aptitude Test (Powers & Swinton, 1981) in which DI item sets helped to define a varimax factor called "data interpretation and technical comprehension" along with items from technical reading passages and items from the 1977 version of the analytical ability measure that seemed similar to the technical reading passages in content and style. In a factor analysis (Rock, Werts, & Grandy, 1982) that involved intercorrelations of item-type part scores paralleling those employed in this study, the loading of the DI items on the quantitative factor was less than the loadings for QC and RM items.

The uniquely high predictive validity of regular mathematics items for mathematics majors, and evidence of differential validity for QC and RM items across fields, suggest the potential for improved assessment in separate consideration of the quantitative item types.

With respect to the analytical ability measure, perhaps the most intriguing aspect of the findings that have been reviewed is (a) the rather persistent indication that AR items tend to exhibit "quantitative" characteristics while LR items tend to exhibit "verbal" characteristics, and (b) that LR items may tend to be more valid than AR items. Powers and Swinton (1981) found that logical reasoning items included in the 1977 version of the analytical ability measure were highly related to a reading comprehension factor. And, with regard to the comparative validity of the LR and AR item types, Wild, Swinton, and Wallmark (1982, Table 22) reported that a subtest containing a 74 percent to 26 percent mix of 19 analytical and logical reasoning items was less closely related to SR-UGPA than a subtest of the same length than included <u>only</u> logical reasoning items (for example, $r = .204$ for the AR/LR combination vs SR-UGPA as compared to $r = .269$ for a 19-item subtest including only logical reasoning items). Combining these two item types in a single score would appear to blunt their predictive effectiveness (see Table 9 and related discussion); moreover, the findings raise questions regarding the desirability of including more AR than LR items in the analytical ability measure since the logical reasoning items appear to have greater criterion-related validity than the analytical reasoning items.

The results that have been reviewed point up the value of evidence regarding the criterion-related validity of item types within the more

general verbal, quantitative, and analytical ability measures. Such evidence could be helpful (a) as a factor to be considered in determining the mix of items in a given ability measure--for example, in decisions regarding the proportional mix of existing item types or decisions to add or eliminate particular item types and (b) in assessments of construct validity--for example, as supplementary to the findings of factor analysis. Using data available in GRE files it would be feasible to develop, and update periodically, basic correlational results for all fields based on pooled departmental data for samples of enrolled undergraduates.*

*Previous studies employing SR-UGPA as an external academic criterion (for example, Miller & Wild, 1979; Wild, Swinton, & Wallmark, 1982; Goodison & Wild, 1982) have been based on total correlation matrices (that is, test-UGPA correlations were computed for samples that were heterogeneous with respect to undergraduate department, even though homogeneous with respect to, say, broad graduate major areas). The direction and extent of covariation among means of departments on the GRE and SR-UGPA variables are not predictable--differences in mean SR-UGPA by department cannot be assumed to reflect differences in level of undergraduate performance. Accordingly, the interpretation of analyses based on total correlation matrices is complicated by the fact that such matrices include the theoretically unpredictable among-means covariances as well as the within-department covariances (see Appendic C).

References

Educational Testing Service. (1981). GRE 1981-82 Information Bulletin. Princeton, N.J.: Author

Goodison, M., & Wild, C. (1982, September). Evaluation of the Graduate Record Examinations (GRE) General (Aptitude) Test, 1981-82. Unpublished manuscript, Educational Testing Service, Princeton, N.J.

Linn, R. L., Harnisch, D. L., & Dunbar, S. B. (1981). Validity generalization and situational specificity: An analysis of the prediction of first-year grades in law school. Applied Psychological Measurement, 5, 281-289.

Miller, R., & Wild, C. L. (Eds.). (1979). Restructuring the Graduate Record Examinations Aptitude Test (GRE Technical Report). Princeton, N.J.: Educational Testing Service.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. Journal of Applied Psychology, 65, 373-406.

Powers, D. E., Swinton, S. S., & Carlson, A. B. (1977) A factor analytical study of the GRE Aptitude Test (GRE Board Professional Report 75-11P). Princeton, N.J.: Educational Testing Service.

Powers, D. E., & Swinton, S. S. (1981) Extending the measurement of graduate admission abilities beyond the verbal and quantitative domains. Applied Psychological Measurement, 5, 141-158.

Ramist, L. (1981a, February 12). Validity of the SAT-verbal subscores. Internal memorandum, Educational Testing Service, Princeton, N.J.

Ramist, L. (1981b, July 7). Further investigation of the validity of SAT-verbal subscores. Internal memorandum, Educational Testing Service, Princeton, N.J.

Rock, D. A., Werts, C., & Grandy, J. (1982). Construct validity of the GRE Aptitude Test across populations--An empirical confirmatory study (GRE Board Professional Report 78-1P & ETS RR 81-57). Princeton, N.J.: Educational Testing Service.

Schrader, W. B. (1984). Three studies of SAT-Verbal item types (College Board Report No. 84-7 & ETS RR 84-33). New York: College Entrance Examination Board.

Wallmark, M. (1982a). Aptitude Test Form 3DGR1 (SR-82-35). Unpublished statistical report, Educational Testing Service, Princeton, N.J.

Wallmark, M. (1982b). Aptitude Test Form 3DRG2 (SR-82-23). Unpublished statistical report, Educational Testing Service, Princeton, N.J.

Wild, C. L., Swinton, S. S., & Wallmark, M. (1982). Research leading to the revision of the format of the Graduate Record Examinations Aptitude Test in October 1981 (GRE Board Professional Report 80-1bP & ETS RR 82-55). Princeton, N.J.: Educational Testing Service.

Willingham, W. W. (1974). Predicting success in graduate education. Science, 183, 273-278.

Wilson, K. M. (1974). The contribution of measures of aptitude and achievement in predicting college grades (ETS RB 74-36). Princeton, N.J.: Educational Testing Service.

Wilson, K. M. (1979). The validation of GRE scores as predictors of first-year performance in graduate study: Report of the GRE Cooperative Validity Studies Project (GRE Board Report 75-8R). Princeton, N. J.: Educational Testing Service.

Wilson, K. M. (1982). A study of the validity of the restructured GRE Aptitude Test for predicting first-year performance in graduate study (GRE Board Report 78-6R). Princeton, N.J.: Educational Testing Service.

Appendix A

Supplementary Data on GRE Part Scores

Six forms of the GRE General Test were administered between October 1981 and September 1982. Table A shows (a) the number of examinees in the sample taking each form and their distribution by sex and (b) means and standard deviations of scores on selected test variables, namely, the verbal, quantitative, and analytical scaled scores, equated across forms, and the raw scores on the various item-type subtests. The latter are not equated across forms, and the average difficulty level of the items making up each subtest may vary within tests for a given form as well as across tests.

Based on the GRE scaled total scores, examinees who took forms used in the first three administrations were somewhat more able than those who took the three forms used in the last two administrations. Males constituted a majority of examinees taking certain forms, while females constituted a majority of examinees taking other forms (for example, in April and June). A majority of all members of the study sample were female.

It may be determined that the part-score means do not covary consistently with the scaled total score means, although a tendency toward positive covariation across forms between raw part scores and total scaled scores is discernible. Data not tabled indicated that the raw total scores on the respective tests covaried closely with the total scaled scores.

In z-scaling all raw scores by test form, using means and standard deviations for all examinees taking each form regardless of administration date, it was assumed (a) that there would be attenuating effects on the relatinship of the z-scaled scores to SR-UGPA associated with lack of equating, but (b) that those effects would be random with respect to item types across forms, and thus (c) that the relative weighting of particular item types would not be influenced by any systematic biasing effect. Evidence suggesting that these assumptions were generally valid is provided in Appendix B.

Table A. Means and Standard Deviations of Raw Part and Converted Total Scores for Selected GRE General Test Takers During 1981-82, by Test Form and Administration Dates

Form	3DGR1	3DGR2	3DGR3	3EGR1	K-3DGR3	3EGR2	All Forms
Males	1584	1314	449	921	170	150	4588
Females	1497	1295	392	1116	206	209	4715
Total*	3103	2633	846	2055	378	360	9375
Admins	Oct-Dec	Oct-Dec-Feb	Oct-Feb	Apr	Apr	June	Total
VARIABLE	MEAN	MEAN	MEAN	MEAN	MEAN	MEAN	MEAN
ANTONYMS	12.3139	12.7372	12.9291	10.6409	11.8571	11.1889	12.0599
ANALOGY	10.0190	10.5063	11.3901	11.1684	10.6492	9.4167	10.5334
SENT.COM	9.4218	9.6934	10.2896	9.7299	9.6005	9.7333	9.6069
READING	16.5756	14.4603	14.6619	13.5173	13.6825	14.7611	14.9521
VOCAB.	22.3329	23.2434	24.3191	21.8092	22.4974	20.6056	22.5934
RD. COMP	25.9974	23.9537	24.9515	23.2472	23.2831	24.4944	24.5590
QUANT. C	21.3974	21.9271	21.9397	20.8642	20.3968	22.5778	21.4832
REG. MT.	13.0922	12.5268	15.7199	12.9294	14.5661	12.9083	13.1872
DATA INT	7.4170	6.7281	6.1797	6.6526	5.4550	5.9028	6.8070
ANAL. R.	23.8263	24.0122	22.2624	21.9504	21.0185	20.5250	23.0862
LOG. R	7.0184	6.5097	7.4764	6.3786	6.8730	7.4556	6.7875
GRE-V	528.5305	517.2541	538.6998	499.2895	503.5185	505.3889	517.4744
GRE-Q	588.6626	579.4873	605.3428	559.1436	558.7566	561.5278	578.8725
GRE-A	574.6955	569.3619	585.1655	555.2360	546.0582	534.0833	567.9307
VARIABLE	SIGMA(N)	SIGMA(N)	SIGMA(N)	SIGMA(N)	SIGMA(N)	SIGMA(N)	SIGMA(N)
ANTONYMS	4.0589	3.9728	3.4459	3.8423	3.5747	3.7212	3.9913
ANALOGY	3.1742	2.9821	2.8632	3.4235	3.0477	2.7119	3.1711
SENT.COM	2.7632	2.5597	2.3053	2.4778	2.5410	2.8198	2.6114
READING	3.8222	3.9677	3.7918	3.7590	3.9661	3.9064	4.0344
VOCAB.	6.6689	6.2688	5.6577	6.6132	6.0195	5.6955	6.4519
RD. COMP	6.0059	5.9404	5.3778	5.6495	5.9882	6.1169	5.9651
QUANT. C	4.4306	5.0260	4.1294	4.6956	4.9110	5.0409	4.7081
REG. MT.	3.4681	3.4646	3.3028	4.1464	3.8896	3.9044	3.7549
DATA INT	1.8758	2.0520	2.2562	2.0196	2.3490	1.9204	2.0422
ANAL. R.	6.2379	7.1577	5.7825	6.3161	6.3834	6.4944	6.5942
LOG. R	2.3949	2.4659	2.3148	2.3740	2.4667	2.3952	2.4345
GRE-V	114.2644	112.4384	106.4413	112.0335	114.0155	108.8927	113.1023
GRE-Q	124.3398	127.4319	114.5938	129.6506	135.7218	130.6996	127.1092
GRE-A	120.1158	121.9327	116.5365	121.2936	125.9737	120.7879	121.2685

*Includes individuals not identifiable by sex

BEST COPY AVAILABLE

Appendix B

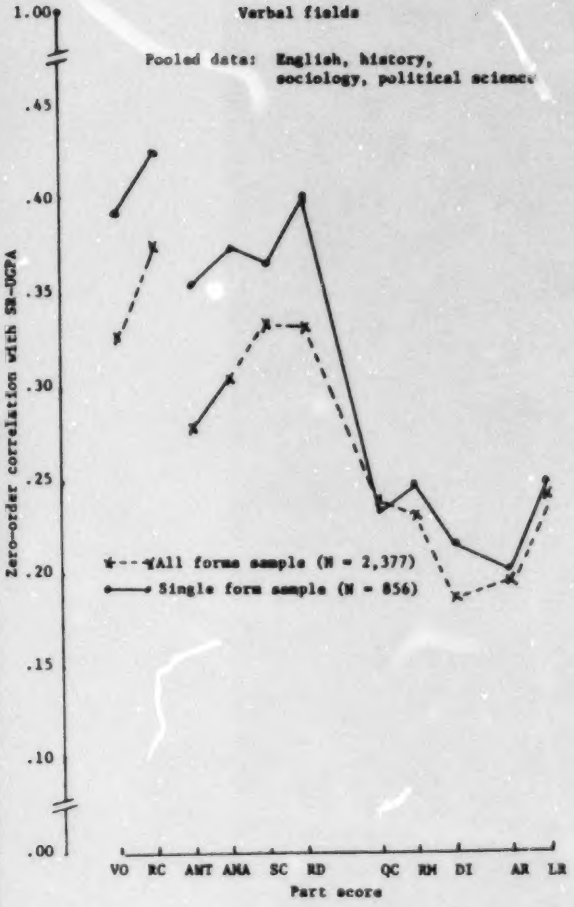
Comparability of Part-Score Validity Profiles for Single Form and Multiple Form Unequated Score Samples

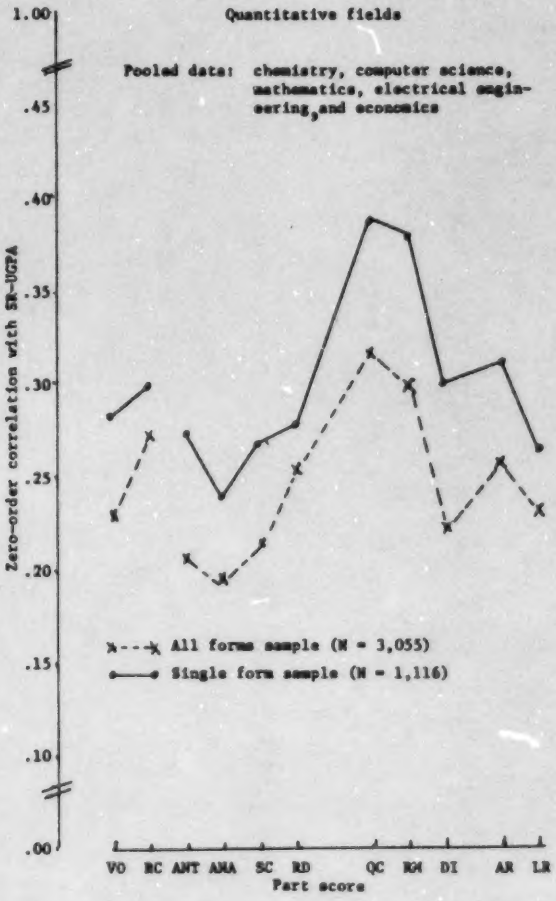
Regression analyses using unequated total verbal, quantitative, and analytical scores from six different forms of the General Test and analyses employing the three GRE scaled total scores, respectively yielded entirely comparable results (see text, Table 14, and related discussion). The relative weighting of the three total scores was consistent across analyses. As expected, the level of correlation was higher for the equated total scores than for the unequated total scores, due, it is assumed, to errors associated with lack of equating across test forms.

Parallel analyses employing equated and unequated part scores were not feasible. However, intercorrelation matrices were generated for examinees taking a single form of the test. The pattern of correlations of part scores with SR-UGPA in this sample may be compared with that for examinees taking several forms, with unequated scores, by reference to Figure B-1.

The part-score correlational profiles for the single-form and multiple-form samples are quite similar, but the level of test-criterion correlations tends to be higher in the single-form sample than in the multiple-form sample. These results suggest strongly that conclusions regarding the relative criterion-related validity of various item-type part scores, based on the findings of the present study employing unequated scores, would be applicable for equated part scores.

Figure B-1. Profiles of pooled within-department correlations of GRE part scores with SR-UGPA for examinees in verbal and quantitative fields, respectively, (a) who took a single form of the GRE General Test and (b) who took up to six different test forms, scores not equated





Appendix C

Factors Involved in the Use of Total vs Pooled Within-Group Correlations in Validation Research

All regression analyses in this study employed pooled within-department correlation matrices. All variables were z-scaled within each department before pooling. Other research employing the self-reported UGPA as an academic criterion (for example, Miller & Wild, 1979; Wild, Swinton, & Wallmark, 1982; Goodison & Wild, 1982) has used total correlation matrices in which coefficients were based on data for all individuals in departmentally heterogeneous samples.

Such total sample correlations are difficult to interpret because it cannot be assumed that differences in mean GPA across several departments represent substantive differences in achievement--the nature of the among means correlation between GRE scores and GPA across several departments is theoretically unpredictable since it is influenced by arbitrary differences in grading standards among departments.

Exhibit C.1 provides scatterplots of GRE verbal (or quantitative) means and first-year graduate GPA means for samples of students from graduate departments that participated in a study of the 1977 restructured GRE General Test (Wilson, 1982).

- Overall, the scatterplot of GRE quantitative and GPA means for chemistry, computer science, economics, and mathematics departments (upper portion of the exhibit) suggests a low positive correlation among departmental means.

- In the lower portion of the exhibit, it may be seen that, among education departments, there is a clear tendency for mean graduate GPA to vary <u>inversely</u> with mean GRE score, while, for the English departments, the scatter of means suggests a generally positive, curvilinear relationship.

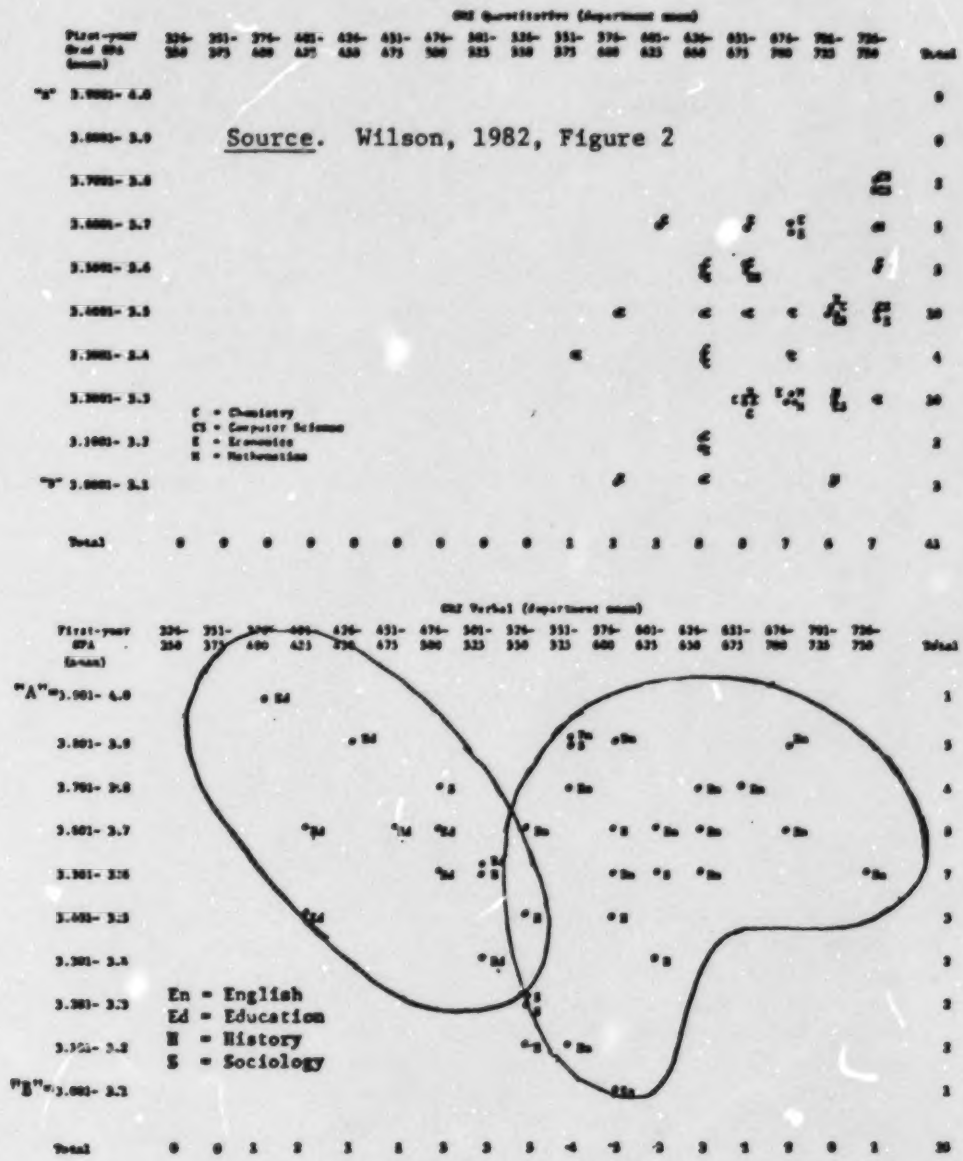
The trends illustrated in the exhibit are consistent with the proposition that neither the degree not the direction of covaration between departmental GRE and GPA means can be assumed to follow a predictable pattern. Moreover, it is reasonable to infer that the <u>total</u> GRE-GPA correlations for education and English majors would differ even though the pooled within-group (within-department) correlations were identical. If such were the case, the total GRE-GPA correlation should be higher for the English sample (with positive among-means correlation) than for the education sample (with negative among-means correlation).

Using data from the present study <u>total</u> correlations between SR-UGPA and the respective GRE item-type pass scores (prior to within-department standardization) were computed to provide a basis for comparison with the pooled within-department correlations actually used in the study. Illustrative findings are summarized in Figure C.1. Note, for example, that

BEST COPY AVAILABLE

Exhibit C.1

Covariation between mean GRE score and mean graduate GPA for selected departmental samples

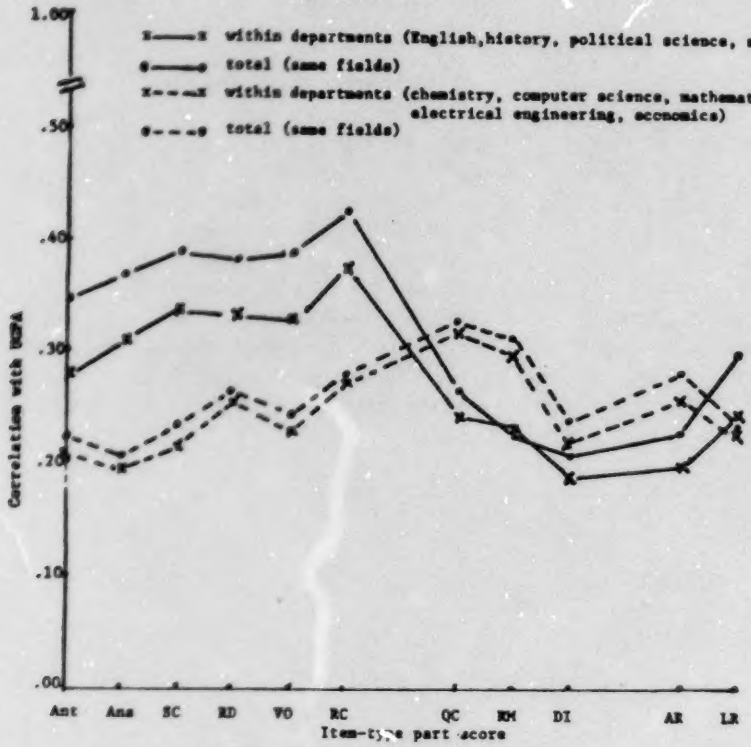


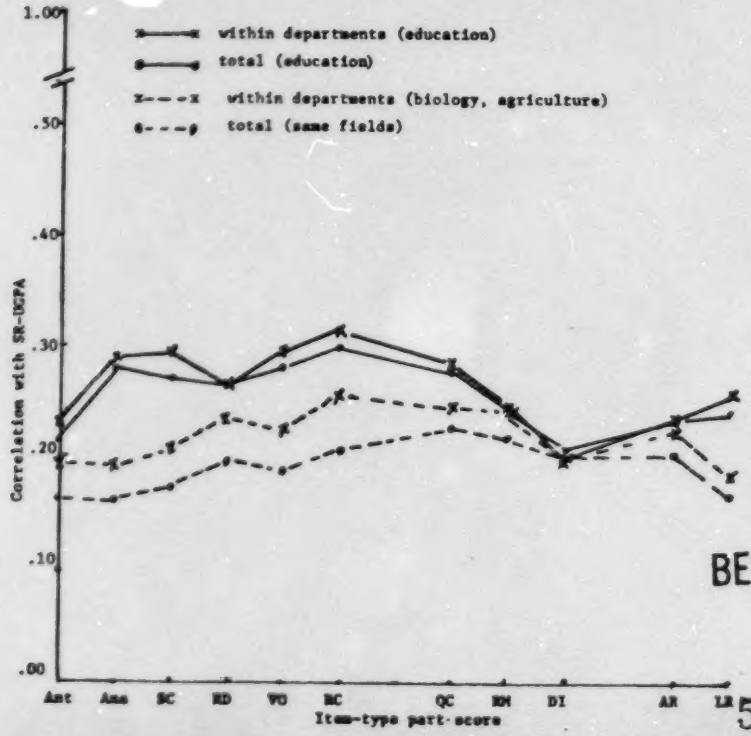
Source. Wilson, 1982, Figure 2

Mean GRE Aptitude Test score (GRE-V or GRE-Q as appropriate to a field) in relation to mean Year 1 graduate GPA for 35 departmental samples from primarily verbal fields and 41 samples from primarily quantitative fields.

BEST COPY AVAILABLE

Figure C.1. Comparison of pooled within-department and total correlation between SR-UGPA and GRE scores in selected fields





BEST COPY AVAILABLE

in the data for education, and the combined agriculture and biology samples, the total correlation is systematically lower than the pooled within-department correlation while the opposite tends to be true for the verbal sample.

The use of total rather than within-group correlations in the present study probably would have led to somewhat different outcomes. Conclusions regarding the relative level of validity of particular subtests for various disciplines would have been affected, for example. It is not clear whether or how outcomes bearing on the relative contribution of the various subtests to prediction of the SR-UGPA criterion might have been affected.

Strictly speaking, it would seem that the most rigorously designed studies of GRE correlations with SR-UGPA would call for the use of pooled, within-department matrices. In validation research involving GPA criteria, the use of total correlations in departmentally heterogenous samples involves elements of interpretive ambiguity that can be avoided only by using pooled within-group correlations.